Probabilistic Simultaneous Pose and Non-Rigid Shape Recovery

Francesc Moreno-Noguer Josep M. Porta
Institut de Robòtica i Informàtica Industrial (CSIC-UPC)
08028, Barcelona, Spain

Abstract

We present an algorithm to simultaneously recover non-rigid shape and camera poses from point correspondences between a reference shape and a sequence of input images. The key novel contribution of our approach is in bringing the tools of the probabilistic SLAM methodology from a rigid to a deformable domain. Under the assumption that the shape may be represented as a weighted sum of deformation modes, we show that the problem of estimating the modal weights along with the camera poses, may be probabilistically formulated as a maximum a posterior estimate and solved using an iterative least squares optimization.

An extensive evaluation on synthetic and real data, shows that our approach has several significant advantages over current approaches, such as performing robustly under large amounts of noise and outliers, and neither requiring to track points over the whole sequence nor initializations close from the ground truth solution.

1. Introduction

Recovering non-rigid 3D shape from monocular sequences is known to be a highly ambiguous problem because very different shapes may have a similar projection [9, 19]. As shown in Fig. 1 the problem becomes even further ill-conditioned if the camera moves while the shape deforms, and both non-rigid shape and camera motion have to be simultaneously retrieved. In order to turn this problem into a tractable one, it is required to introduce prior knowledge about the object's behavior or camera properties.

Standard approaches to limit the space of possible solutions involve introducing deformation models, either physically inspired ones [5, 13, 14] or learned from training data [3, 6, 10, 12, 17, 27]. Surface deformations are then expressed as weighted sums of modes, and estimating shape amounts to retrieving the weights of this linear combination by minimizing and image based objective function. However, convergence is only guaranteed if both the camera pose and shape are initialized close to their true values.

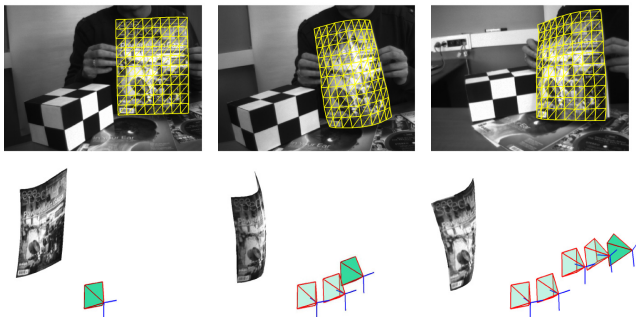

Figure 1. Simultaneous estimation of non-rigid shape and camera pose from input images. **Top:** Three different frames of an input sequence with the recovered mesh overlaid. **Bottom:** Re-textured side view of the retrieved surface and sample camera poses up to the current frame. Note that from only the observation of the deforming shape, estimating the camera pose is very complicated even for the human eye. It would be much easier from the observation of the rigid objects, such as the calibration box, although we do not contemplate this case in the current paper.

Recently, non-rigid structure-from-motion (NRSFM) methods [1, 18, 20, 23, 25] have shown that deformation modes can be learned along with the shape and motion parameters. While these approaches are especially interesting in situations where training data is hard to obtain, they typically require a number of points to be tracked throughout the whole sequence, which is difficult to satisfy in practice. Furthermore, existing NRSFM approaches have shown to be effective only for relatively small deformations and they are quite sensitive to the presence of outliers and noisy data.

In this paper, we propose a new formulation to the problem of simultaneously retrieving non-rigid shape and camera motion that overcomes some of the limitations of previous approaches. We make two basic assumptions, shared with many state-of-the-art approaches [16, 19, 21, 22]. First, we assume that the deformation modes are available, and second, that 2D-to-3D correspondences can be established with a reference image in which the shape is known. Yet, in contrast to NRSFM methods, we do not require tracking the points along the whole sequence, that is,

This work has been partially funded by the Spanish Ministry of Science and Innovation under projects DPI2008-06022 and Consolider Ingenio 2010 CSD2007-00018; and by the EU project GARNICS FP7-247947.

each input image may have its independent set of matches. In addition, our approach tolerates significant rates of outliers and noise. Taking our inspiration on a recent work on SLAM for mapping rigid and static environments [8], we show that, under the above assumptions, the problem of estimating the modal weights describing the shapes and the pose parameters of the camera can be probabilistically formulated as a maximum a posterior estimate that can be iteratively solved using linearization and an efficient QR factorization for sparse linear systems [7]. As we will demonstrate through testing on both synthetic and real data, besides the robustness to outliers and noise, our method does not require from a precise initialization, which is a marked step-forward when compared to the approaches mentioned above used to fit deformation modes to image sequences.

2. Related Work

Many solutions have been proposed over the years that make use of prior information to disambiguate the problem of 3D reconstruction of non-rigid surfaces from monocular images. These solutions may be roughly classified in those based on deformation modes and those that learn the modes along with the shape and pose parameters.

The former include approaches that use physically-inspired modes such as superquadrics [14], thin-plate splines [13] or balloons [5] to reduce the degrees of freedom of the problem. Yet, all these early approaches are only effective to capture relatively small deformations. More realistic deformations were described by complex non-linear models [2, 24], although their applicability is limited to very specific materials.

This limitation is addressed by methods that learn the deformation modes from training data, such as the Active Appearance and Shape Models [6, 12] or the 3D Morphable Models [3]. These approaches represent surface deformations as linear combinations of rigid modes, and retrieving shape entails minimizing an image-based objective function. However, since this function is typically highly non-convex, it requires from good pose and shape initializations to converge, which makes these methods appropriate for tracking shapes with a small inter-frame deformation, such faces [17, 27]. In [10], a similar approach is used to detect human shape and pose from just a single image, although it requires manual pose initialization.

Several recent methods have been proposed to recover non-rigid shape from single images, by using deformation modes in conjunction with local rigidity constraints to reconstruct inextensible surfaces [9, 19, 21, 22], and in conjunction with shading constraints to reconstruct stretchable surfaces [16]. However, none of these approaches retrieves the camera pose, and either assume that the deformation modes are aligned with the camera referencial or yield a solution shape for which the pose is unknown.

In contrast to previous approaches, non-rigid structure-from-motion methods [1, 4, 18, 20, 23, 25, 26] do not require to know a priori the deformation modes and, given a video sequence, compute them simultaneously with the pose and shape. This generality, though, comes at the price of having to impose several constraints that are difficult to hold in practice, such as requiring a sufficient number of points to be tracked throughout the whole sequence. In addition, these methods have only been effectively used to retrieve relatively small deformations, and tend to be sensitive to noisy correspondences, missing data, and outliers.

To address the limitations of previous approaches, we propose a solution inspired in the SLAM technique developed in Robotics to simultaneously recover a set of camera poses and landmarks position in a rigid environment [8]. In this paper we show that using deformation modes in a similar formulation yields an efficient solution to the problem of simultaneously retrieving non-rigid shape and pose. Furthermore, this solution is shown to have significant advantages in terms of robustness and convergence properties.

3. Simultaneous Pose and Non-Rigid Shape

Based on a similar formalism as [8] used for rigid environments, we next show that the problem of simultaneously recovering pose and non-rigid shape can be probabilistically formulated as a maximum a posterior estimate, where both pose and shape are maximized given a set of 3D-to-2D correspondences between each input image and a reference image. We then show that the solution can be iteratively approximated solving a sequence of linear least squares problems.

3.1. Notation and Initial Assumptions

We represent the surface as a triangulated 3D mesh with n_v vertices $\mathbf{v}_i$ concatenated in a vector $\mathbf{x}=[\mathbf{v}_1^\top,\dots,\mathbf{v}_{n_v}^\top]^\top$, and the camera pose as a 6-dimensional vector $\boldsymbol{\rho}$ parameterizing a rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$. Given a sequence of input images $\mathcal{I}=\{\mathbf{I}^k\}$, $1\leq k\leq n_I$, our goal is to simultaneously estimate the surface shape $\mathbf{x}^k$ and camera pose $\boldsymbol{\rho}^k$ at each time instant k.

We assume that we are given a set of 3D points $\mathcal{R}^{ref}=\{\mathbf{r}_i\}$, $1\leq i\leq n_r$ on a *reference configuration* $\mathbf{x}^{ref}$, and that for each input image, we know $n_c^k\leq n_r$ 3D-to-2D correspondences between a subset of points of $\mathcal{R}^{ref}$ and a set of 2D points $\mathcal{U}^k=\{\mathbf{u}_i^k\}$ on $\mathbf{I}^k$. Note that this subset of correspondences is independent at each time step, which relieves of having to track points throughout the whole sequence, as required in NRSFM approaches.

Additionally, we model surface deformations as linear combinations of a mean shape $\mathbf{x}_0$ and n_m deformation modes $\mathbf{Q}=[\mathbf{q}_1,\dots,\mathbf{q}_{n_m}]$

$$\mathbf{x}^k=\mathbf{x}_0+\sum_{i=1}^{n_m}\alpha_i^k\mathbf{q}_i=\mathbf{x}_0+\mathbf{Q}\boldsymbol{\alpha}^k\,, \tag{1}$$

where $\boldsymbol{\alpha}^k = [\alpha_1^k, \ldots, \alpha_{n_m}^k]^\top$ are unknown weights that define the surface shape at time k. In our implementation, the deformation modes were obtained by applying Principal Component Analysis over a training database of meshes with similar deformations as the target motion.

And finally, we also assume the camera to be calibrated and denote by $\mathbf{A}$ its matrix of intrinsic parameters.

3.2. Probabilistic Formulation of the Problem

By representing the surface shape through deformation modes, we can reformulate our problem as that of estimating $\Phi = \{\boldsymbol{\phi}^k\}$ with $1 \leq k \leq n_I$ and with $\boldsymbol{\phi}^k = (\boldsymbol{\rho}^{k\top}, \boldsymbol{\alpha}^{k\top})^\top$ the augmented $(6 + n_m)$-state vector that collects the unknown pose and shape for a given frame k, given the observations $\mathcal{U} = \{\mathcal{U}^k\}$, and where $\mathcal{U}^k = \{\mathbf{u}_i^k\}$, $1 \leq i \leq n_c^k$ are the known measurements of the 2D coordinates of the n_c^k 3D-to-2D correspondences at time instant k. This can be expressed in terms of the following maximum a posterior estimate

$$\Phi^* = \arg\max_{\Phi} P(\Phi \mid \mathcal{U}) \propto \arg\max_{\Phi} P(\Phi, \mathcal{U}) \; . \quad (2)$$

The joint probability $P(\Phi, \mathcal{U})$ may be written as

$$P(\Phi, \mathcal{U}) = P(\boldsymbol{\phi}^1) \prod_{k=2}^{n_I} P(\boldsymbol{\phi}^k | \boldsymbol{\phi}^{k-1}) \prod_{i=1}^{n_c^k} P(\mathbf{u}_i^k | \boldsymbol{\phi}^k, \mathbf{r}_{i_k}) \quad (3)$$

where $P(\boldsymbol{\phi}^1)$ is a prior on the initial pose and shape, $P(\boldsymbol{\phi}^k | \boldsymbol{\phi}^{k-1})$ is the motion model and $P(\mathbf{u}_i^k | \boldsymbol{\phi}^k, \mathbf{r}_{i_k})$ is the measurement model of the 3D reference point $\mathbf{r}_{i_k}$, corresponding the i-th 3D-to-2D match at time step k.

3.3. Motion and Measurement Models

We describe the motion model on the camera pose and modal weights in the form $\boldsymbol{\phi}^k = f(\boldsymbol{\phi}^{k-1}) + \mathbf{w}_{\boldsymbol{\phi}}^k$, which may be probabilistically written as

$$P(\boldsymbol{\phi}^k | \boldsymbol{\phi}^{k-1}) \propto \exp -||f(\boldsymbol{\phi}^{k-1}) - \boldsymbol{\phi}^k||^2_{\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^k} \quad (4)$$

where $||\cdot||^2_{\boldsymbol{\Sigma}}$ denotes the squared Mahalanobis distance, $f(\cdot)$ models the camera and shape dynamics, and $\mathbf{w}_{\boldsymbol{\phi}}^k$ is a zero mean Gaussian noise with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^k$. In fact, this covariance is a block diagonal matrix, composed of a 6×6 covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\rho}}^k$ for the poses, and a $n_m \times n_m$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^k$ for the modal weights. As shown in Fig 2, we set these covariance matrices to relatively large values in order to produce many different types of shapes and poses. This increases the generality of our approach to solve problems where input data may considerably differ from the training data we used to compute the modes.

The measurement model is described in the form $\mathbf{u}_i^k = h(\boldsymbol{\phi}^k, \mathbf{r}_{i_k}) + \mathbf{w}_{\mathbf{u}}^k$, and hence

$$P(\mathbf{u}_i^k | \boldsymbol{\phi}^k, \mathbf{r}_{i_k}) \propto \exp -||h(\boldsymbol{\phi}^k, \mathbf{r}_{i_k}) - \mathbf{u}_i^k||^2_{\boldsymbol{\Sigma}_{\mathbf{u}}^k} \quad (5)$$

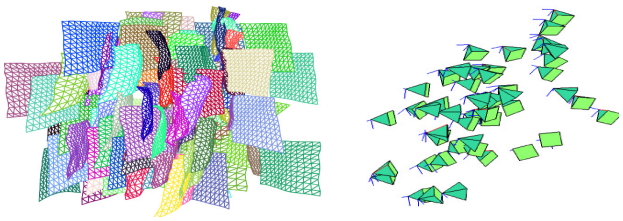

Figure 2. Shape and Pose priors samples we consider, obtained by adding Gaussian noise to a given shape and pose. Note that the priors we use are fairly ambiguous and allow representing many different configurations of shapes and poses.

where $\mathbf{u}_i^k$ is the known 2D correspondence of $\mathbf{r}_{i_k}$, $h(\cdot)$ is the measurement equation, and $\mathbf{w}_{\mathbf{u}}^k$ is a zero mean Gaussian noise with 2×2 covariance matrix $\boldsymbol{\Sigma}_{\mathbf{u}}^k$. The function $h(\boldsymbol{\phi}^k, \mathbf{r}_{i_k})$, corresponds to the equation that projects the 3D reference point $\mathbf{r}_{i_k}$ onto the image, after being mapped according to the pose and shape parameters $\boldsymbol{\rho}^k$ and $\boldsymbol{\alpha}^k$, respectively.

More specifically, let $\mathbf{p}_i^k$ be a point on the mesh $\mathbf{x}^k$ corresponding to the point $\mathbf{r}_{i_k}$ in the reference configuration. We can write $\mathbf{p}_i^k$ in terms of the barycentic coordinates of the face it belongs as

$$\mathbf{p}_i^k = \sum_{j=1}^{3} a_{ij} \mathbf{v}_{i_j}^k \; , \quad (6)$$

where the a_{ij} are the barycentric coordinates and $\mathbf{v}_{i_j}^k$ are the vertices of the face in $\mathbf{x}^k$ containing the point $\mathbf{p}_i^k$. Since we assume the mesh does not stretch, these barycentric coordinates remain constant for each point an can be easily computed from points $\mathbf{r}_{i_k}$ and the reference mesh $\mathbf{x}^{ref}$.

The measurement equation $h(\boldsymbol{\phi}^k, \mathbf{r}_{i_k})$ returns $\tilde{\mathbf{u}}_i^k$, the 2D projection of $\mathbf{p}_i^k$ onto the image given the current pose and shape parameters. If we expand $\boldsymbol{\phi}^k$ into a rotation matrix $\mathbf{R}^k$, translation vector $\mathbf{t}^k$, and modal weights $\boldsymbol{\alpha}^k$ we can write such a projection as

$$w_i \begin{bmatrix} \tilde{\mathbf{u}}_i^k \\ 1 \end{bmatrix} = \mathbf{A} \begin{bmatrix} \mathbf{R}^k | \mathbf{t}^k \end{bmatrix} \begin{bmatrix} \mathbf{p}_i^k \\ 1 \end{bmatrix}$$

where w_i is a scalar. Finally, by injecting the barycentric coordinates of Eq. 6 and the modal description of Eq. 1, the measurement equation $h(\boldsymbol{\phi}^k, \mathbf{r}_{i_k}) = \tilde{\mathbf{u}}_i^k$ can be written in terms of the pose parameters and modal weights.

3.4. Least Squares Formulation

As discussed in Section 3.2, the problem of simultaneously retrieving pose and shape can be formulated in terms of a MAP estimate of the joint probability $P(\Phi, \mathcal{U})$. By taking the negative logarithm of Eq. 2, and considering Eqs. 3, 4 and 5, it can be further shown that the problem may be reduced to the following non-linear least-squares

estimation

$$\Phi^* = \arg\min_{\Phi} \sum_{k=1}^{n_I} \varepsilon(\boldsymbol{\phi}^{k-1}, \boldsymbol{\phi}^k) \tag{7}$$

where

$$\varepsilon(\boldsymbol{\phi}^{k-1}, \boldsymbol{\phi}^k) = \left\| f(\boldsymbol{\phi}^{k-1}) - \boldsymbol{\phi}^k \right\|^2_{\boldsymbol{\Sigma}^k_{\boldsymbol{\phi}}} + \sum_{i=1}^{n_c^k} \left\| h(\boldsymbol{\phi}^k, \mathbf{r}_{i_k}) - \mathbf{u}_i^k \right\|^2_{\boldsymbol{\Sigma}^k_{\mathbf{u}}}$$

and where we define $f(\boldsymbol{\phi}^0) = \mathbf{0}$.

Since the measurement function $h(\cdot)$ is nonlinear and the process function $f(\cdot)$ may also be non-linear, the minimum is iteratively approximated linearizing Eq. 7. Let $\boldsymbol{\theta}_0 = (\boldsymbol{\phi}_0^{1\top}, \ldots, \boldsymbol{\phi}_0^{n_I\top})^\top$ be an initial estimation of Φ^*. We approximate $f(\cdot)$ and $h(\cdot)$ linearizing at $\boldsymbol{\theta}_0$

$$\begin{aligned} f(\boldsymbol{\phi}^{k-1}) &\approx f(\boldsymbol{\phi}_0^{k-1}) + \mathbf{F}^{k-1}\boldsymbol{\delta}^{k-1} \\ h(\boldsymbol{\phi}^k, \mathbf{r}_{i_k}) &\approx h(\boldsymbol{\phi}_0^k, \mathbf{r}_{i_k}) + \mathbf{H}_{i_k}^k \boldsymbol{\delta}^k \end{aligned}$$

where $\boldsymbol{\delta}^k = \boldsymbol{\phi}_0^k - \boldsymbol{\phi}^k$, $\mathbf{F}^{k-1}$ is the $(6+n_m)\times(6+n_m)$ Jacobian of $f(\cdot)$, and $\mathbf{H}_{i_k}^k$ is the $2\times(6+n_m)$ Jacobian matrix of $h(\cdot)$, both of them evaluated at the corresponding element of $\boldsymbol{\theta}_0$

$$\mathbf{F}^{k-1} = \left.\frac{\partial f(\boldsymbol{\phi}^{k-1})}{\partial \boldsymbol{\phi}^{k-1}}\right|_{\boldsymbol{\phi}_0^{k-1}} \qquad \mathbf{H}_{i_k}^k = \left.\frac{\partial h(\boldsymbol{\phi}^k, \mathbf{r}_{i_k})}{\partial \boldsymbol{\phi}^k}\right|_{\boldsymbol{\phi}_0^k}$$

Let us write the error in the dynamic motion prediction as $\mathbf{c}^k = \boldsymbol{\phi}_0^k - f(\boldsymbol{\phi}_0^{k-1})$, and the error in the measurement as $\mathbf{d}_i^k = \mathbf{u}_i^k - h(\boldsymbol{\phi}_0^k, \mathbf{r}_{i_k})$. Then, Eq. 7 becomes

$$\varepsilon(\boldsymbol{\delta}^{k-1}, \boldsymbol{\delta}^k) \approx \left\| \mathbf{F}^{k-1}\boldsymbol{\delta}^{k-1} + \mathbf{G}\boldsymbol{\delta}^k - \mathbf{c}^k \right\|^2_{\boldsymbol{\Sigma}^k_{\boldsymbol{\phi}}} + \sum_{i=1}^{n_c^k} \left\| \mathbf{H}_{i_k}^k \boldsymbol{\delta}^k - \mathbf{d}_i^k \right\|^2_{\boldsymbol{\Sigma}^k_{\mathbf{u}}}$$

where $\mathbf{G}$ is a $(6+n_m)\times(6+n_m)$ identity matrix, introduced to simplify subsequent notation. Finally, the original least-squares problem is re-written as

$$\boldsymbol{\delta}^* = \arg\min_{\boldsymbol{\delta}} \|\mathbf{B}\,\boldsymbol{\delta} - \mathbf{b}\|^2_{\boldsymbol{\Sigma}} \tag{8}$$

where $\boldsymbol{\delta} = [\boldsymbol{\delta}^{1\top}, \ldots, \boldsymbol{\delta}^{n_I\top}]^\top$, and $\boldsymbol{\Sigma}$ is a matrix made of all the $\boldsymbol{\Sigma}^k_{\boldsymbol{\phi}}$ and $\boldsymbol{\Sigma}^k_{\mathbf{u}}$ noise terms. The matrix $\mathbf{B}$ collects all Jacobian matrices, and the vector $\mathbf{b}$ all errors in motion prediction and measurements

$$\mathbf{B} = \begin{bmatrix} \mathbf{G} & & & & & \\ \mathbf{F}^1 & \mathbf{G} & & & & \\ & \mathbf{F}^2 & \mathbf{G} & & & \\ & & \ddots & \ddots & & \\ & & & \mathbf{F}^{n_I-1} & \mathbf{G} & \\ \mathbf{J}^1 & & & & & \\ & \mathbf{J}^2 & & & & \\ & & & \ddots & & \\ & & & & & \mathbf{J}^{n_I} \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} \mathbf{c}^1 \\ \mathbf{c}^2 \\ \mathbf{c}^3 \\ \vdots \\ \mathbf{c}^{n_I} \\ \mathbf{d}^1 \\ \mathbf{d}^2 \\ \vdots \\ \mathbf{d}^{n_I} \end{bmatrix}$$

with $\mathbf{J}^k = [\mathbf{H}_1^k, \cdots, \mathbf{H}_{n_c^k}^k]^\top$ and $\mathbf{d}^k = (\mathbf{d}_1^k, \ldots, \mathbf{d}_{n_c^k}^k)$.

In order to solve Eq. 8 we derive and equate to zero. Then, $\boldsymbol{\delta}^*$ may be found as the solution of

$$\hat{\mathbf{B}}^\top\,\hat{\mathbf{B}}\,\boldsymbol{\delta}^* = \hat{\mathbf{B}}^\top\,\hat{\mathbf{b}}$$

with $\hat{\mathbf{B}} = \boldsymbol{\Sigma}^{-1/2}\,\mathbf{B}$ and $\hat{\mathbf{b}} = \boldsymbol{\Sigma}^{-1/2}\,\mathbf{b}$. Note that $\mathbf{B}$ is a large but very sparse matrix. Assuming a constant number n_c of 3D-to-2D correspondences per image, $\mathbf{B}$ would be a $n_I(n_m{+}6{+}2n_c)\times n_I(n_m{+}6)$ matrix. Typical values in our experiments are $n_I = 50$ images, we detect about $n_c = 100$ correspondences per image, and use $n_m = 30$ deformation modes. These values would yield a 11800×1800 matrix, although with only a 2.3% of non-null entries. Taking advantage of this sparsity, $\hat{\mathbf{B}}$ and $\hat{\mathbf{b}}$ can be directly defined premultiplying $\mathbf{F}^k$, $\mathbf{G}$, and $\mathbf{c}_i^k$ by $(\boldsymbol{\Sigma}^k_{\boldsymbol{\phi}})^{-1/2}$, and $\mathbf{H}_i^k$ and $\mathbf{d}_i^k$ by $(\boldsymbol{\Sigma}^k_{\mathbf{u}})^{-1/2}$. The resulting $\hat{\mathbf{B}}$ is also sparse and we can use a sparse QR factorization [7] on $\hat{\mathbf{B}}$ to efficiently obtain $\boldsymbol{\delta}^*$ without explicitly having to compute $\hat{\mathbf{B}}^\top\,\hat{\mathbf{B}}$. The solution $\boldsymbol{\delta}^*$ is then used to update $\boldsymbol{\theta}_0$ and the procedure described in this Section is repeated until convergence.

3.5. Detecting and Removing Outliers

As will be shown in the results section, the formulation we propose, yields very good results in terms of convergence and robustness to noise. In addition we implemented a procedure similar to what was proposed in [21] to detect and remove 3D-to-2D correspondences with very large errors. We define

$$\lambda_i^k = \frac{\mathbf{d}_i^k}{\text{median}(\mathbf{d}_i^k\,, 1 \le k \le n_I\,, 1 \le i \le n_c^k)}\,,$$

and reduce the influence of the more noisy correspondences, by multiplying the rows of $\mathbf{B}$ with the weight

$$w_i^k = \begin{cases} \exp(-\lambda_i^k) & \text{if } \lambda_i^k < \lambda \\ 1 & \text{otherwise} \end{cases}$$

where the parameter λ is chosen large enough (we set $\lambda = 3$ in all our experiments) to ensure that only those measurements with large errors $\mathbf{d}_i^k$ are penalized. Yet, we initially do not remove these observations, because their gross error might come from a wrong estimate of shape and pose at the current iteration. Instead, we remove them if after having contributed in the current estimate, their reprojection error remains outside a radius, that is reduced at each iteration. In practice, we start with a 100 pixel radius that progressively reduce until a value of 10 pixels.

4. Experimental Results

In this section we extensively evaluate the performance of our approach against noise in the correspondences, the presence of outliers, or its dependence on the quality of the

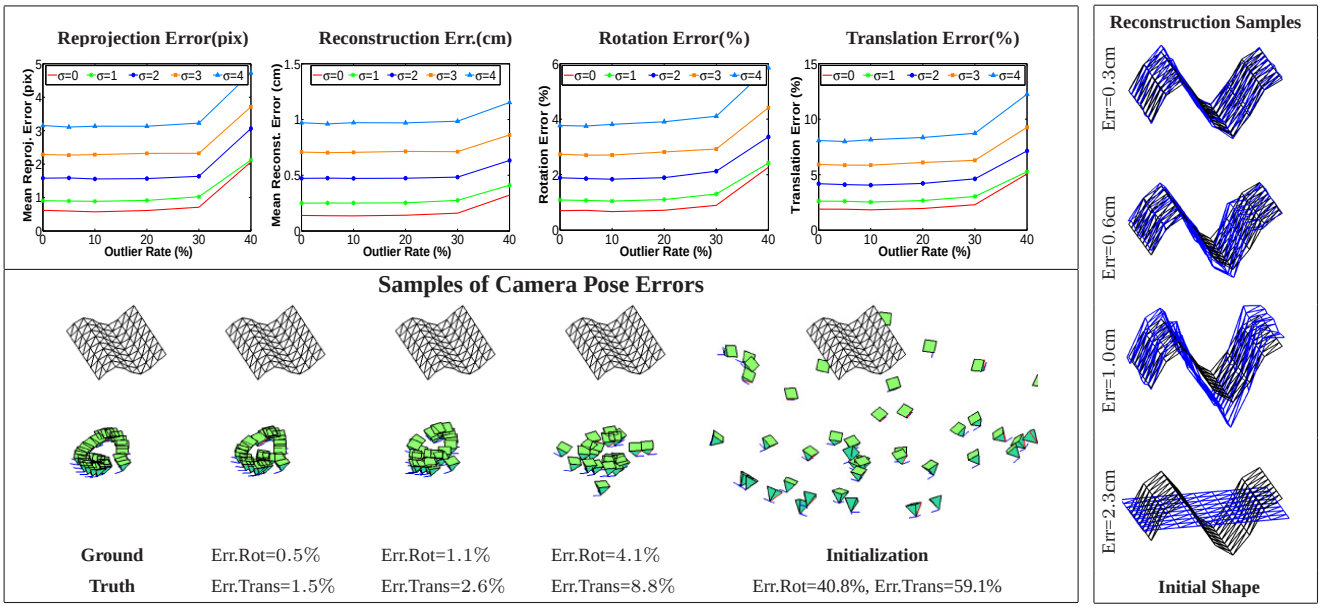

Figure 3. Results on shape and pose recovery for a sequence of a deforming synthetic mesh. **Top-Left:** Mean reprojection, reconstruction, rotation and translations errors, as a function of the number of outliers and for different levels of noise in the correspondences. **Bottom-Left, Right:** Different levels of pose and reprojection errors to give significance to the errors we obtain. Note, that even with fairly vague initializations, our algorithm converges to reasonably good solutions.

initialization. We show results on both synthetic and real images.

In order to position the current approach within the state-of-the-art, we also provide a comparison against [18], which is representative of the non-rigid shape from motion techniques. Although the two methods are not directly comparable, as they require from different assumptions, we enforce the comparison and show the benefits of using combined priors on the shape deformation and camera dynamics, even when they are very weak.

4.1. Synthetic Data

We first applied our approach to a 50 frames synthetic sequence of a 9×9 mesh, simulating the deformation of a wave with increasing amplitude. The reference configuration was represented by a 30×30 cm planar shape. The camera was allowed to move according to random Brownian paths on the surface of a 80 cm sphere centered on the mesh, and with the optical axis pointing to this center. The left-most graph in the middle row of Fig. 3 shows one example of shape and camera poses generated this way.

For each pair of camera pose and mesh shape, we then synthetically produced 150 random 3D-to-2D correspondences, between a 640×480 image acquired for that particular shape and pose, and the reference configuration. Given this setup, we performed two different types of experiments, to evaluate both the robustness and convergence performance of the proposed algorithm.

In the first experiment we analyzed the performance of our approach against noise in the 2D correspondences and the presence of outliers. More specifically, we performed 10 different experiments by adding noise with standard deviation of $\{0, 1, 2, 3, 4\}$ pixels, and by introducing a percentage of outliers of $\{0, 5, 10, 20, 30, 40\}\%$. In addition, this combination of parameters was repeated for 15 different random camera paths.

In each of these experiments all the shapes in the sequence were initialized with the planar shape of the reference mesh. The poses, were initialized by adding random noise to the ground truth poses such that the initial percentage of rotation and translations errors were approximately of 50%. The right-most graphs in the second and third rows of Fig. 3 show that these initializations are significantly far from the ground truth solutions.

In all experiments we used the same set of parameters to describe the dynamic and measurement models of Section 3.3. As a dynamic model, we used simple Brownian motion, and the function $f(\cdot)$ in Eq. 4 was taken to be the identity, that is, $f(\boldsymbol{\phi}) = \boldsymbol{\phi}$. The covariance matrix $\boldsymbol{\Sigma}_\rho$ on the poses was set to a constant diagonal matrix, with a 0.1 standard deviation for the rotation components, and 3 cm for the translational ones. The covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$ on the modal weights was computed from the training data used to estimate the deformation modes, scaled by a factor of 3 to increase the generality of the method. For the measurement model we set a diagonal covariance $\boldsymbol{\Sigma}_{\mathbf{u}}$ to the observations, with a 3 pixel standard deviation.

Fig. 3 reports the mean results of the experiment. In the

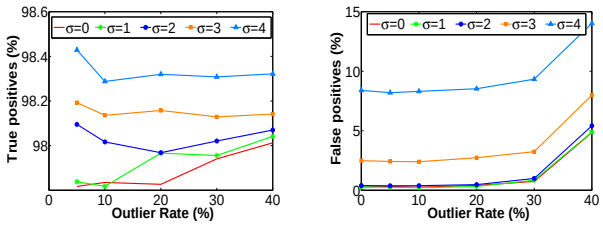

Figure 4. Detection of outliers. P: true number of outliers. N: true number of inliers. TP: Number of outliers correctly detected. FP: Number of inliers miss-classified as outliers. True Positives(%)=$\frac{TP}{P}$. False Positives(%)=$\frac{FP}{N}$. Observe that even for large levels of noise, our algorithm correctly detects most of the outliers.

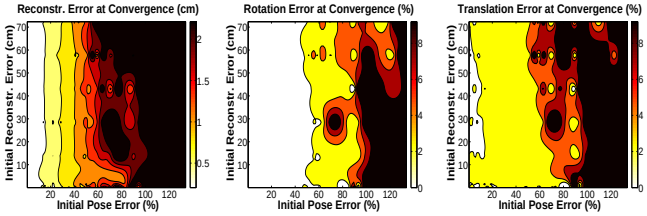

Figure 5. Shape and pose errors at convergence, as a function of the error in the initialization.

upper row, we plot the mean reprojection, reconstruction and pose errors, as a function of the percentage of outliers and for different levels of noise in the correspondences. Observe that even for large levels of noise and outliers, the results are within reasonable bounds. The images on the middle and bottom rows give significance of the errors we obtain. For instance, observe that a mean reconstruction error of 1.0 cm, still represents a good approximation to the true shape.

In Fig. 4 we evaluate the methodology described in Section 3.5 to detect outliers. Observe that it yields very large rates of true positives and low rates of false positives. This means that correctly detects most of the outlier correspondences, while only miss-classifies a very small percentage of correct correspondences. Of course, the results slightly fall when the noise in the correspondences is increased, because then, correct but very noisy correspondences are classified as outliers.

In a second experiment with the synthetic data we evaluated the convergence behavior of our approach. For that purpose, we initialized our algorithm with very different poses and shapes, either relatively close to the true solutions or very far away. Fig. 5 shows the reconstruction and pose errors at convergence as a function of the errors in the initialization. Errors larger than specific ratios are saturated and shown in black, such that we can consider the black regions, as non-convergence areas. In fact these non-convergence values are reasonable values for which the retrieved solutions are visually disturbing. Observe that convergence almost does not depend on the quality of the initial shapes, and the initial pose is the dominant factor. That being said, our algorithm tolerates errors in the initial pose of up to 80%, which is relatively large, specially considering that the pose error shown in the middle-right plot of Fig. 3 is of about 50%.

4.2. Real Images

We tested our method on a 120-frames sequence of a bending paper and a 100-frames sequence of a deforming T-shirt, acquired with a Pointgrey Bumblebee stereo camera. In both cases the camera was moved around the deforming shape while capturing the sequence. The upper images in Fig. 1 show three different frames of the "bending paper" sequence, where the movement of the camera can be clearly appreciated from the viewpoint change of the still calibration box.

We used the stereo rig to estimate the ground truth shape, although our algorithm was ran with just the images from one of the cameras. The ground truth pose was computed by applying the PnP algorithm [15] over a small set of manually introduced correspondences between points on a 3D model of the calibration box and points in each of the input images. The 3D-to-2D correspondences between the reference configuration and the input images of the mesh were computed using SIFT [11]. In order to obtain a sufficiently large number of correspondences, we used the SIFT matches to initially detect the surface in 2D and then used normalized cross correlation in very small regions to obtain dense correspondences.

Since the distance between the camera and the surface was roughly the same as for the synthetic experiments, and the inter-frame camera displacement was also very similar we used the same dynamic and measurement models we defined in the previous section.

Yet, an issue we had to resolve was that the number of frames of the real sequences was much larger than for the synthetic case, and the size of the matrix $\mathbf{B}$ in Eq. 9 became very large to be tractable. To handle this situation we implemented an *incremental* version of our algorithm, in which the sequence was split into several parts, and each part solved independently. However, in order to avoid jumps between the different parts, we allowed certain overlapping of the frames and shared their solution among sub-sequences.

Fig. 6 depicts the results for the two real experiments. In each case, the upper-row graphs plot the errors per frame, at initialization, after 4 iterations, and at convergence. Since the results have been obtained by applying our algorithm to several sub-sequences the number of iterations to converge is not unique. However, all sub-sequences converged using between $50 - 70$ iterations. The bottom plots, show the configuration of camera poses at previous time instances.

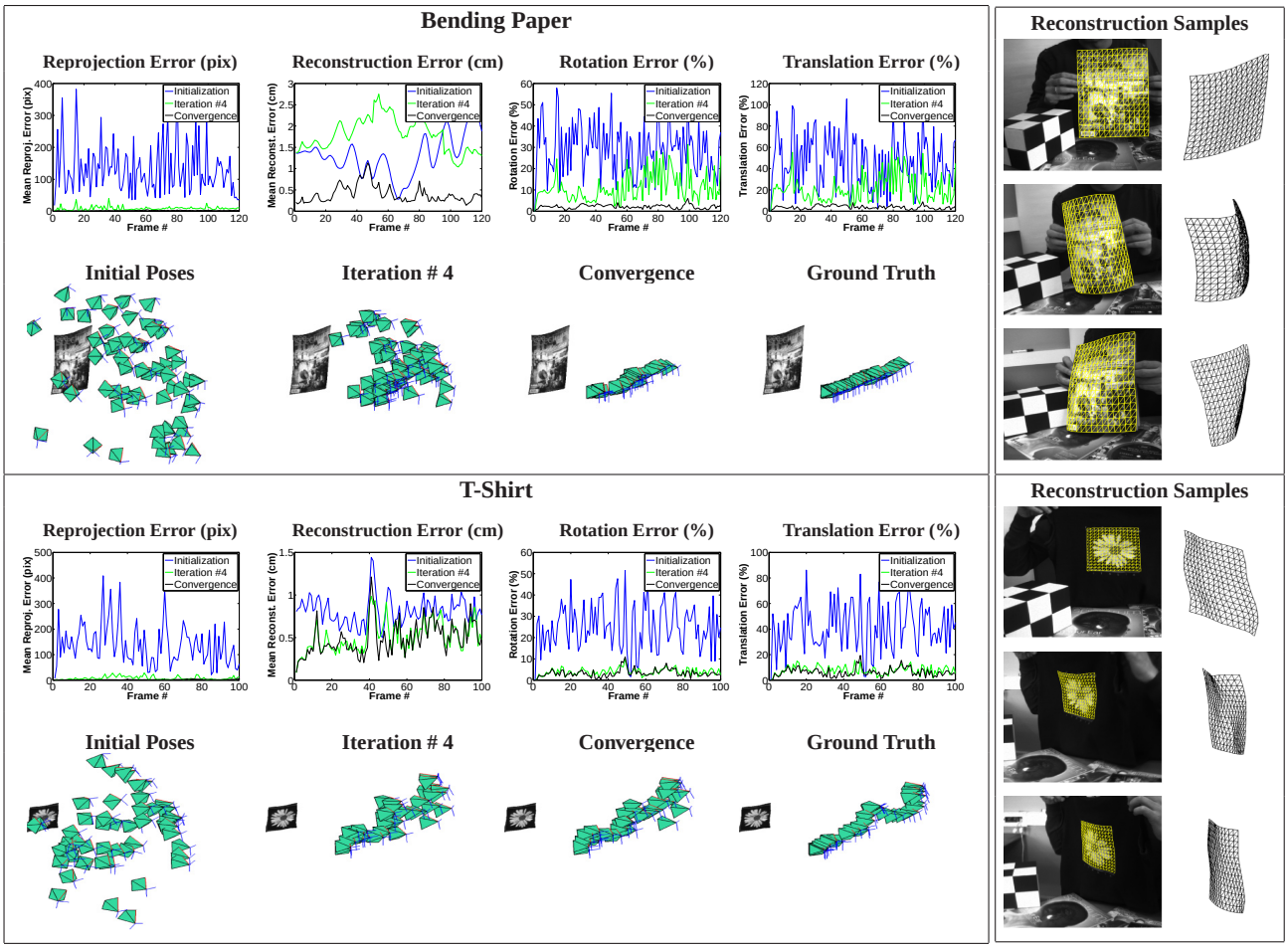

Figure 6. Results on two real sequences. **Left:** For each experiment, the upper plots depict the errors per frame obtained after initialization, 4 iterations and convergence. The bottom plots show the poses corresponding to the previous time instances. **Right:** Shape recovery on two real sequences. For each frame, the we overlaid the recovered mesh overlaid on the original image, and show the 3D mesh seen from a constant point of view, after eliminating the camera movement.

Note, that our algorithm yields fairly good results, specially considering the large error of the initial set of poses. Finally, in Fig. 6-Right we show the recovered shape for different frames of each sequence.

4.3. Comparison with NRSFM techniques

We finally compare our method against [18], a recent Non-Rigid Shape From Motion algorithm. As said above, there are substantial differences between our approach and NRSFM methods. The most important is that we make use of a deformation model, computed from training data, while in NRSFM methods do not assume that training data is available, and simultaneously estimate 3D shape and modes. This obviously makes these algorithms more general, but at the price of being more sensitive to noise and constrained to relatively small deformations.

Fig. 7 shows the results of the comparison for the synthetic sequence used in Section 4.1. In order to satisfy the input data requirements of [18], we provided the tracks of all the vertices of the mesh, and projected them onto the image using an orthographic camera model. We then computed the reconstruction error for increasing levels of input noise. As expected, the behavior of the NRSFM methods is quite poor, and becomes specially unstable for large amounts of noise. In contrast, the use of the deformation modes yields a remarkable robustness and stability to our algorithm. In addition, besides shape, our algorithm also provides an accurate estimation of the camera pose, which we did not show here because [18] does not explicitly compute pose.

5. Conclusion

In this paper we have shown that the problem of simultaneously retrieving pose and non-rigid shape given a set of 3D-to-2D correspondences can be probabilistically formulated as a maximum a posteriori (MAP) estimate. We then introduce dynamic and measurement models accounting for noisy data, and reduce the MAP estimate to a non-

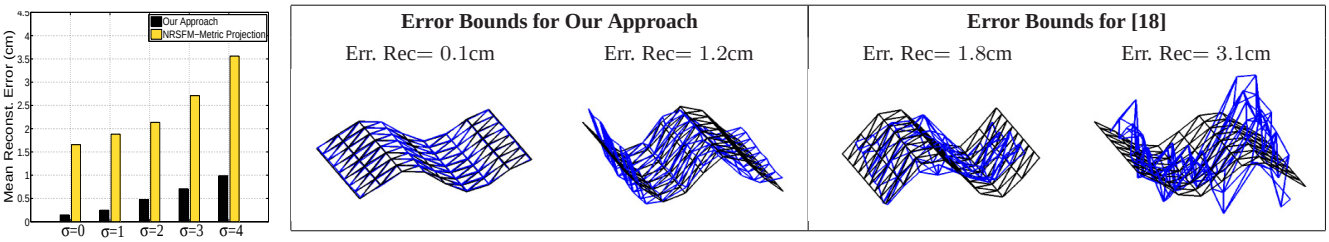

Figure 7. Comparison with NRSFM aproaches. **Left:** Reconstruction error of our approach and [18] for the synthetic sequence presented in Sect. 4.1, as a function of the input noise. **Right:** Sample reconstructions showing the error bounds for both methods. Observe, that even the solution with largest error of our approach represents a better approximation than the solution with smallest error obtained with [18]. Of course, this additional accuracy is consequence of using known deformation modes.

linear least squares optimization that we solve using standard techniques. In the results section, we have shown that we obtain satisfactory results under situations where current methods are prone to fail, such as, large rates of outliers and noise in the input data, or very poor quality of the initializations.

The formulation of the problem we propose is very general, and allows introducing additional constraints either on the structure of the mesh or on the dynamic models. In particular, as part of future work, we pretend to investigate the use of length constraints on the edges of the mesh, as has been done in previous literature [19, 22]. We believe that introducing these constraints would reduce the dependence of the method on training data, and that more global deformation modes might be used for several applications.

References

[1] A. Bartoli, V. Gay-Bellile, U. Castellani, J. Peyras, S. Olsen, and P. Sayd. Coarse-to-fine low-rank structure-from-motion. In *CVPR*, 2008.

[2] K. S. Bhat, C. D. Twigg, J. K. Hodgins, P. K. Khosla, Z. Popovic, and S. M. Seitz. Estimating cloth simulation parameters from video. In *Eurographics Symposium on Computer Animation*, 2003.

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3–d faces. In *SIGGRAPH*, 1999.

[4] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *CVPR*, 2000.

[5] L. Cohen and I. Cohen. Finite-element methods for active contour models and balloons for 2-d and 3-d images. *PAMI*, 1993.

[6] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *ECCV*, 1998.

[7] T. Davis. Multifrontal multithreaded rank-revealing sparse qr factorization (submitted). 2009.

[8] F. Dellaert and M. Kaess. Square root sam: Simultaneous localization and mapping via square root information smoothing. *IJRR*, 2006.

[9] A. Ecker, A. D. Jepson, and K. N. Kutulakos. Semidefinite programming heuristics for surface reconstruction ambiguities. In *ECCV*, 2008.

[10] P. Guan, A. Weiss, A. Balan, and M.J.Black. Estimating human shape and pose from a single image. In *ICCV*, 2009.

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[12] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 2004.

[13] T. McInerney and D. Terzopoulos. A finite element model for 3D shape reconstruction and nonrigid motion tracking. In *ICCV*, 1993.

[14] D. Metaxas and D. Terzopoulos. Constrained deformable superquadrics and nonrigid motion tracking. *PAMI*,1993.

[15] F. Moreno-Noguer, V. Lepetit, and P. Fua. Accurate non-iterative O(n) solution to the PnP problem. In *ICCV*, 2007.

[16] F. Moreno-Noguer, M.Salzmann, V. Lepetit, and P. Fua. Capturing 3D stretchable surfaces from single images in closed form. In *CVPR*, 2009.

[17] E. Muñoz, J. Buenaposada, and L. Baumela. A direct approach for efficiently tracking with 3D morphable models. In *ICCV*, 2009.

[18] M. Paladini, A. D. Bue, M. Stosic, M. Dodig, J. Xavier, and L. Agapito. Factorization for non-rigid and articulated structure using metric. In *CVPR*, 2009.

[19] M. Perriollat, R. Hartley, and A. Bartoli. Monocular template-based reconstruction of inextensible surfaces. In *BMVC*, 2008.

[20] V. Rabaud and S. Belongie. Linear embeddings in non-rigid structure from motion. In *CVPR*, 2009.

[21] M. Salzmann and P. Fua. Reconstructing sharply folding surfaces: A convex formulation. In *CVPR*, 2009.

[22] M. Salzmann, F. Moreno-Noguer, V. Lepetit, and P. Fua. Closed-form solution to non-rigid 3D surface registration. In *ECCV*, 2008.

[23] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *PAMI*, 2008.

[24] L. V. Tsap, D. B. Goldgof, and S. Sarkar. Nonrigid motion analysis based on dynamic refinement of finite element models. *PAMI*, 2000.

[25] R. Vidal and R. Hartley. Perspective nonrigid shape and motion recovery. In *ECCV*, 2008.

[26] J. Xiao and T. Kanade. Uncalibrated perspective reconstruction of deformable structures. In *ICCV*, 2005.

[27] W. Zhang, Q. Wang, and X. Tang. Real time feature based 3D deformable face tracking. In *ECCV*, 2008.